Sentiment Analysis

A Probabilistic Approach

January 30, 2012

Supervised by I. Langbroek (Blauw Research)
Mentored by dr. M. van Someren (Universiteit van Amsterdam)

S. A. Gieske	S. Laan	C. R. Verschoor	D. S. ten Velthuis	A. J. Wiggers
6167667	6036031	10017321	0577642	6036163

Artificial Intelligence
Faculty of Science
Universiteit van Amsterdam

Contents

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

2 Problem Specification

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

2.1 Data Cleaning

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

2.2 Data Reduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

2.3 Machine Learning

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3 Theory

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.1 Perceptron

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.2 Artificial Neural Network

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.3 Method 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.4 Probabilistic Terms

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.4.1 Accuracy

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.4.2 Precision

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.4.3 Recall

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3.4.4 F-Measure

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

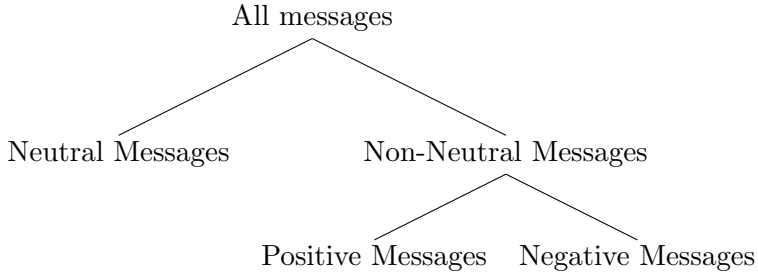
4 Usage and User Guide

5 Implementation and Results

5.1 Global Approach

First we'll explain our general approach. The idea is to use a binary classifier, i.e. either true or false, multiple times. This way we can first decide whether a message is neutral

and then we can decide between the non-neutral messages whether it is a positive or negative message. Maybe a tree structure shows the used approach more clearly:



As you can see, with this approach we need only to distinguish between two classes at a time. This is very convenient, because there are a lot of techniques that can discriminate between two classes, whereas multi-class-classifiers are less common.

5.2 Binary Classification

In this section, we'll describe the method that we use to classify a message. The approach consists of a few steps:

1. The counting of word frequencies
2. Calculation of word probabilities
3. Calculation of message probabilities
4. Finding a good threshold

First we count the frequency of each word, i.e. we count how many times it occurs in the training set. We also keep track of how many times this word was encountered in either class. For example, if we want to decide whether messages are neutral or not, we keep track of how many times each word has been seen in a neutral message. The number of times the word occurs in the opposite class, is equivalent to the total number of encounters, minus the encounters in the first class.

When the frequencies of all words[1] are found, we can calculate a probability for each word, which gives us an idea of how likely it is to be encountered in a certain class. The formula for this probability is the following:

$$P(word) = \frac{\sum word \in C_1}{\sum word \in C_1 \cup C_2} \tag{1}$$

So the probability that a word is in C_1, the first class, is the number of times it has been encountered in a sentence, which was tagged to be in C_1, divided by the total number of encounters.

[1] All words in the training set

Now that all words have probabilities assigned to them, or at least all words in the training set, we can calculate the probabilities of the sentences.

$$P(s) = \frac{1}{n} \sum_{w \in s} P(w) \tag{2}$$

Where s is the sentence, n the number of words in the sentence and w a word in the sentence. As follows from the formula, we take the weighted sum of the word probabilities.

Now that every sentence has an associated probability, the real machine learning can begin. The goal is to find a threshold for the probabilities, i.e. when a sentence probability is higher than a certain amount, we classify it as C_1, otherwise, we classify it as belonging to C_2.

We tried several approaches to accomplish this task.

5.3 Algorithm 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.3 Discussion and Reflection

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum

ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4 Algorithm 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.3 Discussion and Reflection

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

6 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum

ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.